Fast and Expressive Multi-Byte Prediction with Probabilistic Circuits

Andreas Grivas[1], **Lorenzo Loconte**[1], **Emile van Krieken**[2], **Piotr Nawrot**[1], **Yu Zhao**[1], **Euan Wielewski**[3], **Pasquale Minervini**[1,4], **Edoardo Ponti**[1], **Antonio Vergari**[1]

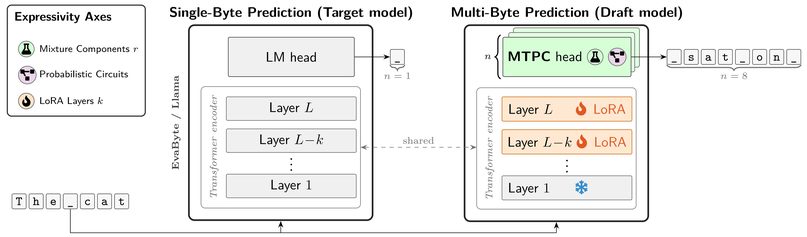
School of Informatics, University of Edinburgh[1] *Vrije Universiteit Amsterdam*[2] *NatWest Group*[3] *Miniml.AI.*[4]

Paper Code Cirkit

1 Multi-Byte Prediction

Goal **Faster text generation...**

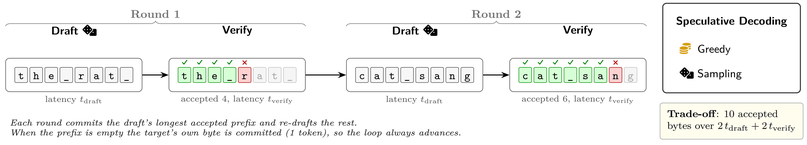
Predict n tokens at each timestep by modelling $q_{\boldsymbol{\theta}}(x_{t+1}, \ldots, x_{t+n} \mid \mathbf{x}_{\leq t})$.



2 Speculative Decoding

Goal **...with no loss in generation quality.**

Reject drafted tokens that do not match the target model (exactly for greedy and in expectation for sampling).



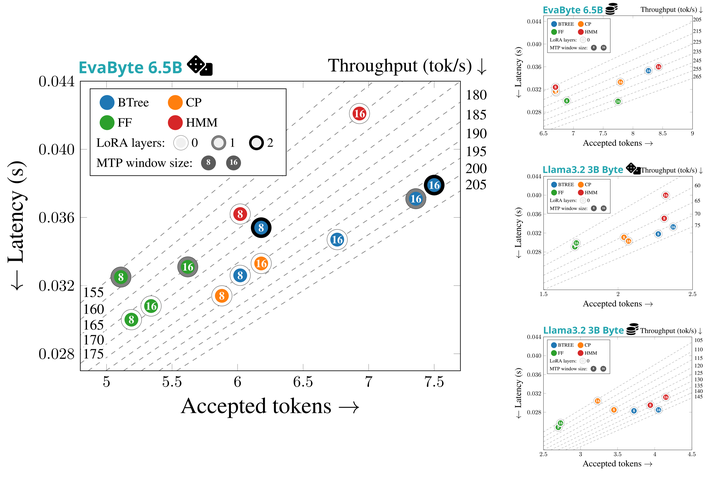
The Expressiveness/Efficiency Trade-off

To increase **throughput** we need to balance **acceptance rate** and **latency**:

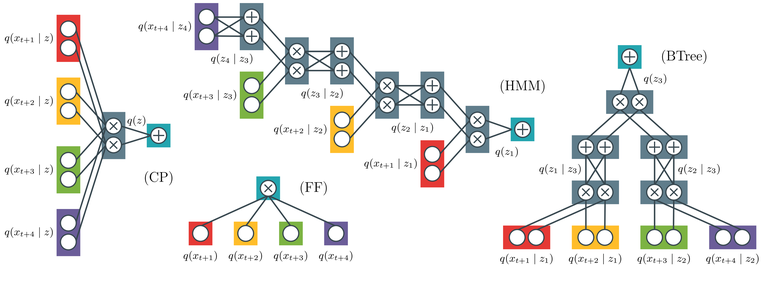
Throughput Optimised Latency Optimised Acceptance Optimised

$$\mu_{\text{tok/s}} \approx \frac{\text{✓}\mu_{\text{acc}}}{\text{⏱}\mu_{\text{lat}}}$$

TL;DR **We speed up a 6.5B byte-level LLM by 5.15× by trading off expressiveness & efficiency with Probabilistic Circuits.**



3 Probabilistic Circuits (PCs)



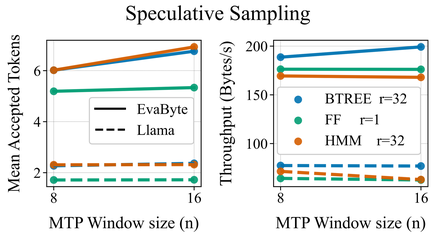
Results (Navigating Expressiveness Axes)

Axis 1 More Mixture Components

PC	r	⏱μ_{lat} ↓	✓μ_{acc} ↑	$\mu_{\text{tok/s}}$ ↑
FF	1	**0.0299**$\pm$0.0003	5.19$\pm$0.05	176.3$\pm$0.9
	8	0.0314$\pm$0.0005	5.67$\pm$0.05	184.2$\pm$3.9
	16	0.0323$\pm$0.0020	5.79$\pm$0.08	183.4$\pm$13.3
CP	32	0.0314$\pm$0.0003	5.88$\pm$0.02	**191.0**$\pm$1.6
	64	0.0327$\pm$0.0015	5.96$\pm$0.04	185.9$\pm$7.7
	128	0.0337$\pm$0.0001	**6.02**$\pm$0.03	182.3$\pm$1.1

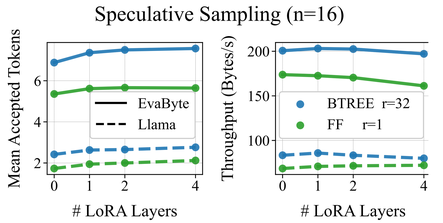
FF CP, $r > 32$ CP, $r = 32$

Axis 2 More Expressive Circuits



HMM, $r = 32$ BTREE, $r = 32$

Axis 3 Draft LoRA Layers



EvaByte BTREE 1 LoRA Layer

Overall Speed-Ups with MTPC

EvaByte 6.5B: $5.15\times$ over Single Byte and $1.17\times$ over FF.
Llama3.2 3B Byte: $2.24\times$ over Single Byte and $1.17\times$ over FF.